DOCUMENT RESUME

ED 065 592 TM 001 861

AUTHOR Edmonston, Leon P.
TITLE A Review of Attempts to Arrive at More Suitable Evaluation Models: An Introspective Look.
PUB DATE Apr 72
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Criterion Referenced Tests; *Decision Making; *Evaluation Methods; *Models; Statistical Analysis
IDENTIFIERS *Southwest Educational Development Lab

ABSTRACT

Attempts by the Southwest Educational Development Laboratory to arrive at a comprehensive evaluation model are reviewed. Problems that arose from using classical procedures and measures are discussed. The emphasis of the Lab was to develop an evaluation model related to criterion referenced measures that provide decision-making information. (DB)

ED 065592

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

A REVIEW OF ATTEMPTS TO ARRIVE AT MORE SUITABLE EVALUATION MODELS: AN INTROSPECTIVE LOOK[1]

Leon P. Edmonston

Southwest Educational Development Laboratory

The procedures employed by the Southwest Educational Development Laboratory (SEDL) in the evaluation of criterion-referenced measures (CRMs) have evolved from the employment of analytical methods which considered only the statistical properties of CRMs to the decision-making model used currently at SEDL (Edmonston, Randall, & Oakland, 1972). These initial attempts at model construction resulted in the rejection of more traditional statistical techniques which failed to address the unique psychometric properties of CRMs. This paper will review these attempts to arrive at a comprehensive evaluation model.

TM 001 861

A previous paper (Randall, 1972) has described differences between criterion-referenced and norm-referenced tests. One distinction maintained between these instruments at SEDL is that CRMs are single, independent items which are sampled theoretically from a large item domain. Performance on these items or "mini-tests" has only to satisfy a binary pass-fail criterion. Employment of standards such as these place serious limitations on the scale properties attributed to the items and thus restrict statistical procedures to associational measures designed for point distributions ($\emptyset$), tests of independence such as χ^2 and its derivatives, and a few other parametric and non-parametric tests.

---

[1] Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1972.

Typically, when a binary scored test item is cross classified with itself or with another variable, results may be arranged in a contingency table. Only the α X α table ($\alpha = 2$) will be discussed in this paper. Consider the following table:

TABLE 1

A FOUR-CELL CLASSIFICATION TABLE
OF PROPORTIONS OF STUDENTS WHO
PASSED OR FAILED EACH OF TWO ITEMS

Item 1 \ Item 2	Pass	Fail	Total
Pass	P11	P12	P1·
Fail	P21	P22	P2·
Total	P·1	P·2	1.0

The most obvious question asked of data displayed in these tables is the degree of association existing between classifications. This is true especially if interest is in reliability and validity estimates. When a "true" dichotomy can be assumed, phi is the appropriate measure for such a test of association.[1] However, the employment of traditional correlational measures such as phi with data obtained from CRMs has proven infeasible at SEDL. This has resulted partly from the fact that the philosophy of student assessment underlying employment of CRMs rejects the notion of intersubject comparisons. Consequently, student variability and the statistical procedures which depend upon it are irrelevant.

---

[1] Measures like the tetrachoric r (r_t) are employed when two variables each have been reduced artificially to two categories. Some disagreement exists as to whether gradations (i.e., degrees of passing and failing) can exist in a true dichotomy; those who assume a continuity hypothesis might employ r_t on the above data. The arguments to be advanced against the use of $\emptyset$ are equally applicable for r_t.

A second reason that the use of phi and other techniques has been rejected can be attributed to the standards of student performance necessary to be maintained in the program. During construction of CRMs, requirements are imposed that items accurately reflect the content specifications of the criterion behavior, that the items are not too difficult for the student population, and that extra-item factors (such as those related to media or instructions to the teachers) will not bias responding. If initial observations also indicate that these requirements have been met and that the objective has been attained satisfactorily[1], the item is submitted for a reliability check. It is in the imposition of such a criterion that traditional measures like phi are rendered inappropriate. For example, with a mean of .8 and a corresponding .16 variance estimate a negatively skewed score distribution results, severely restricting the maximal limit of a positive phi as the pass categories for each variable vary from one another. As an example, 125 first grade Ss were retested within a 10-day interval on five single-item CRMs employed in SEDL's Social Education Program. Results are presented in Table 2.[2]

Obviously, tables A and E are heavily influenced by the zero (0) in the diagonal category; however, the occurrence of empty cells is quite common with CRM data. Moreover, to indicate how phi might differ if the zero cell in tables A and E had frequencies, five Ss were added to the original N (125) and placed in the empty cell in each of the two tables. The phi for table A jumped from -.03 to .57, in table E from -.03 to .51. As evidenced, the estimates are not reliable when the distribution assumes such a skewed form.

---

[1] Satisfactory attainment is defined in terms of a large majority of students, usually at least 80%, reaching criterion.

[2] Table E, Table 2, represents a bad item from the content validation sense in that the heavy performance demand placed upon the student indicated that only 2 Ss passed the initial testing. This item is included here for illustrative purposes.

TABLE 2

RETEST RELIABILITIES (Ø) OBTAINED ON FIVE SINGLE ITEM CRMs EMPLOYED IN SEDL'S SOCIAL EDUCATION PROGRAM

A

Test \ Retest	+	-	
+	.94	.04	.98
-	.02	.00	.02
	.96	.04	1.0

Ø = -.03

B

Test \ Retest	+	-	
+	.70	.10	.80
-	.07	.13	.20
	.77	.23	1.0

Ø = .48

C

Test \ Retest	+	-	
+	.89	.03	.92
-	.06	.02	.08
	.95	.05	1.0

Ø = .31

D

Test \ Retest	+	-	
+	.64	.05	.69
-	.21	.10	.31
	.85	.15	1.0

Ø = .34

E

Test \ Retest	+	-	
+	.00	.02	.02
-	.06	.92	.98
	.06	.94	1.0

Ø = -.03

This is true even though an alternative such as summation of the diagonal frequencies indicates that there is at least 90% agreement between the classifications in those passing or failing on both testings in each table.

The principal problem in these and other examples is due to the unstable and quite restricted estimate of variance. As stressed by Popham and Husek (1969), the very philosophy underlying criterion-referenced testing is the rejection of the relevance of variance. It would naturally follow from this and the experiences related above that there would be a rejection of measures which depend upon variability; however, it is easy to be misled. In tables B, C, and D in Table 2, the phi coefficients are moderately high and a positive relationship exists between polytomies. Results such as these prompted consideration of such questions as how can comparison between several coefficients be made and how can individual coefficients be judged in terms of some external standard.

Concerning comparisons between several phi's, the coefficient of correlation is only a descriptor and not a number on an interval scale. Consequently, such statements that the phi in table C (.31) represents two-thirds the relationship that the phi in table B (.48) indicates are incorrect; unfortunately, statements such as these are often used when communicating results to those involved in making decisions about curriculum objectives and item writing.

Similarly, Fisher's $\bar{Z}$ for evaluating the difference between uncorrelated correlations was rejected because statements pertaining to statistical significance also are unable to provide information concerning the relative difference between the correlations. Eventually, we realized that no adequate intercomparisons could be made between correlations by going the classical statistics route.

Equally important as the comparison between coefficients is the evaluation of the single retest or validity correlation; usually this would connote the

establishment of a standard against which to judge the adequacy of the coefficient. The most obvious approach is to test the null hypothesis that phi is equal to zero and then employ either the standard error of $\emptyset$, $1/\sqrt{N}$, as put forth by McNemar (1962, p. 198) or the χ^2 technique. The whole question thus begins to extend into the employment of significance and the adequacy of this concept for evaluating reliability and validity. For example, if we accept the .01 level, all phi's in tables B, C, and D of Table 2 are significant (df = 1). Actually with N equal to 125, a phi of .23 is significant at the .01 level. This small value is dependent upon the large N; it also indicates that statistical significance is as equally inappropriate for CRMs as it is for NRMs when referring to reliability and validity.

Chi-square and the related $\bar{Z}$ test[1] for the difference between proportions also were attempted. The question asked of the data with these statistics was whether the change in frequencies between the test and the retest was statistically significant. The null hypothesis for a contingency table is that the two variables are independent in the population. However, according to Goodman and Kruskal (1954), because "an excellent test of independence may be based on χ^2, does not at all mean the χ^2, or some simple function of it, is an appropriate measure of degree of association" (p. 740). This has also been emphasized by R. A. Fisher (1948). Chi-square was discarded for reasons such as these.

Although related to χ^2, the $\bar{Z}$ test for correlated proportions was felt to be more appropriate because it tests whether the probability of passing item one is the same as the probability of passing item two when account is taken of the fact that the proportions are correlated. The formula is provided by McNemar (1947):

(1) $$\bar{Z} = \frac{b - c}{\sqrt{b + c}}$$

---

[1] $\chi^2 = Z^2$ with one degree of freedom for both correlated and uncorrelated proportions.

where b and c are equal to the off-diagonal cells, P12 and P21, respectively, in Table 1. In formula (1) b-c also equals the difference between the pass categories for the two polytomies; if the pass categories are equal, b-c equals zero and $\bar{Z}$ equals zero. An example is presented in Table 3.

TABLE 3

CONTINGENCY TABLE OF PROPORTIONS IN WHICH THE $\bar{Z}$ FOR CORRELATED PROPORTIONS IS EQUAL TO ZERO

		Retest: Pass	Retest: Fail	Total
Test	Pass	.50	.10	.60
	Fail	.10	.30	.40
	Total	.60	.40	1.00

$\bar{Z} = .0$

This statistic was intuitively appealing and some time was spent testing it with different values in the off-diagonal cells. We often found, as would be expected from the formula, that when the proportions passing in each of the two classifications was held constant and cell frequencies were varied, there was a good correspondence between $\bar{Z}$ and a summation of the diagonal frequencies, which, when divided by N, Goodman and Kruskal (1954) call the coefficient of agreement. However, there were many instances where $\bar{Z}$ and a diagonal summation failed to fluctuate together. Generally, this occurred when the off-diagonal cells were similar or identical with one another. For example, consider Table 4, tables A, B, and C:

TABLE 4

CONTINGENCY TABLES ILLUSTRATING EXAMPLES OF THE LACK OF AGREEMENT BETWEEN $\bar{Z}$ AND SUMMATION OF DIAGONAL PROPORTIONS ($\Sigma_{\alpha} \rho\alpha\alpha$)

A

Test \ Retest	+	-	
+	.50	.20	.70
-	.20	.10	.30
	.70	.30	1.00

$\bar{Z} = .0$

$\Sigma \rho\alpha\alpha = .60$

B

Test \ Retest	+	-	
+	.40	.10	.50
-	.10	.40	.50
	.50	.50	1.00

$\bar{Z} = .0$

$\Sigma \rho\alpha\alpha = .80$

C

Test \ Retest	+	-	
+	.97	.01	.98
-	.01	.01	.02
	.98	.02	1.00

$\bar{Z} = .0$

$\Sigma \rho\alpha\alpha = .98$

D

Test \ Retest	+	-	
+	.60	.12	.72
-	.08	.20	.28
	.68	.32	1.00

$\bar{Z} = .89$

$\Sigma \rho\alpha\alpha = .80$

E

Test \ Retest	+	-	
+	.20	.60	.80
-	.10	.10	.20
	.30	.70	1.00

$\bar{Z} = 5.98$

$\Sigma \rho\alpha\alpha = .30$

F

Test \ Retest	+	-	
+	.40	.05	.45
-	.25	.30	.55
	.65	.35	1.00

$\bar{Z} = -3.63$

$\Sigma \rho\alpha\alpha = .70$

Considerable variability exists within the diagonals, the proportion in these cells ranging from .60 to .98; $\bar{Z}$, however, equals zero in all instances. From these and other examples (see also tables D, E, and F in Table 4) as well as from the fact that the critical values of $\bar{Z} = 1.65$ and $\bar{Z} = 2.33$ were not meaningful for acceptance or rejection of items as reliable, we realized that it was not the difference between frequencies which should be investigated, but rather the similarities. Also, it was equally obvious from our work with correlational measures that there is no simple relation between measures of association like phi and measures of agreement such as summation of diagonal proportions. Accordingly, we began to rely more heavily upon information within each of the cells of the contingency table without attempting to employ classical measures to summarize them. We also began to employ measures independent of variance. These will be discussed in the next paper.

Problems similar to those already discussed were met with item analysis procedures. Point biserial correlations of items with total scores were rejected immediately for the same reasons as was phi (e.g., as the p values deviate from .50, a very low ceiling is imposed quickly on the maximal correlation[1]). Other item-total techniques, such as comparing item performance of those in the upper and lower 27% of the total score distribution also were deemed as infeasible. This is because the instruments assessing student progress through the curriculum are composed of items measuring heterogeneous abilities and no adequate information could be provided by this measure. The principal emphasis then turned towards reliance upon performance patterns on items measuring the same concept over a series of curriculum unit tests. If

---

[1] Actually, it is not possible to obtain a perfect correlation between a dichotomous and continuous variable unless all scores on the dichotomized variable fall exactly upon two points on the continuous variable, thus making it dichotomous.

N Ss met criterion on all items except the one in question, the conclusion was reached that the item was weak and the curriculum could not be faulted. Conversely, if the same Ss consistently scored lowly on the item under examination as well as upon other items measuring the same concept, the item was considered as discriminating. Unfortunately, this procedure began to assume a rationale related to norm-referenced testing with comparisons being made between Ss who scored highly or lowly on the CRM.

This paper has attempted to illustrate that as SEDL has moved away from reliance upon classical procedures and measures, emphasis has been upon maintenance of a definition of educational evaluation such as advocated by Stufflebeam, et al (1971). They define evaluation as "the process of delineating, obtaining and providing useful information for judging decision alternatives" (p. 40). The results of our quest for an evaluation model related to CRMs which would delineate as well as provide the information appropriate for making judgements about SEDL programs will now be discussed.

REFERENCES

Cox, R. C. & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, February 1966.

Edmonston, L. P., Randall, R. S., & Oakland, T. A model for estimating the reliability and validity of criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1972.

Fisher, R. A. Statistical methods for research workers. New York: Hafner Publishing Co., Tenth Edition, 1948.

Goodman, L. A. & Kruskal, W. H. Measures of association for cross-classifications. American Statistical Association Journal, 1954, 49, 732-764.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 1947, 12, 153-157.

McNemar, Q. Psychological statistics. 3d Ed. New York: Wiley, 1962.

Popham, W. J. & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Randall, R. S. Contrasting norm-referenced and criterion-referenced measures. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1972.

Stufflebeam, D. I., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. I., Merriman, H. O., & Provus, M. M. Educational evaluation and decision making. Peacock Publishers, 1971.